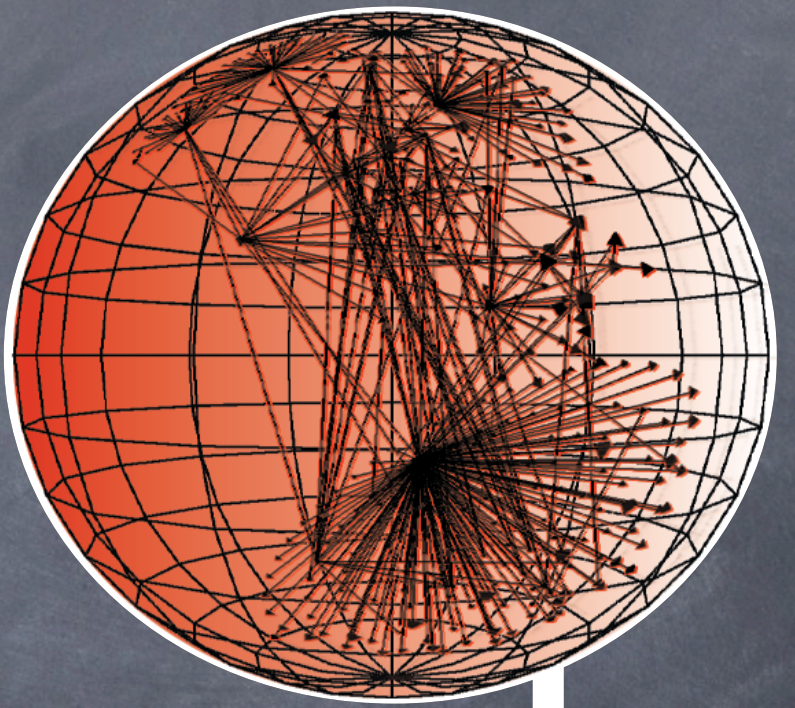
DHS PREDICT project: CAIDA update

Kimberly Claffy, CAIDA

March 21, 2011



caida

DHS PREDICT project: CAIDA update

- Data collection updates
- Dataset dissemination statistics
- Other activities
- Open issues

Data collection - passive

- OC192 backbone: March 2008 - Feb 2011
  - 14 TB compressed, 24.2 TB uncompressed
  - unanonymized: 7 TB compressed, 13.2 TB uncompressed
  - anonymized: 7 TB compressed, 11.0 TB uncompressed
- Changes since December 2010
  - Retired 2007 dataset (was in PREDICT)
  - Completed 2010 Dataset
- Problems:
  - No traces on Chicago monitors in Nov, Dec 2010, Jan 2011 due to hardware failure; fixed in February.
- Plans:
  - 2011 annual dataset in progress (now includes Jan and Feb)
  - strip payload/L1/L2, transfer, anonymize, archive
  - collect 1 hour trace per month = 200-250 GB (compressed)
  - keep a quarterly sample - select the best quality

Data collection - passive

- UCSD telescope:
  - data from most recent 30-days (~five weeks) “live” on disk
    - typically 3 TB compressed, 5 TB uncompressed
  - previous month - backed up on tape (samqfs)
    - current: 2009/12/01 - now
      - 42 TB (compressed), 78 TB (uncompressed)
    - applied for NSF funding
      - analysis
      - develop automated triggers and alerts
      - curate custom data sets upon request
      - explore “near real-time”, “bring code to the data” data sharing
- OC48 traces:
  - 964.5 GB compressed, 1.7 TB
  - unanonymized: 815.7 GB (compressed)
  - anonymized: 148.8 GB (compressed), in PREDICT

Data collection - active

- old skitter data (in PREDICT):
  - 1.47 TB (compressed), 4.02 TB (uncompressed)
    - discontinued February 2008
- current Ark data:
  - IPv4 topology: 1.3 TB (compressed)
  - IPv6 topology: 1.0 GB (compressed)
  - 53 monitors in 30 countries, 16 IPv6 capable
- data curation:
  - create derivative data sets
  - aggregate in ITDK
    - router-level topologies: nodes and links
    - hostnames
    - router-to-AS assignment
    - geographical information
      - http://www.caida.org/data/active/internet-topology-data-kit/
- applied for NSF funding to curate/analyze/annotate IPv6 data

how do we serve the data?

- PREDICT (OC48 traces, topology from skitter, telescope)
- Academics who sign AUP (OC192, topology from Ark, telescope)
- Derived data publicly available i.e., AS-links
- Commercial researchers must join CAIDA
- Aggregated statistics online:
  - topology: http://www.caida.org/projects/ark/statistics
  - traffic: http://www.caida.org/data/realtime/

Meta-data for packet traces

- OC192 data: 2008-2009, Jan-Oct 2010
  - an hour-long trace every month
  - usually, 3rd Thursday, 13:00 - 14:00 UTC
- OC48 data: 2002-2003
- Publicly available statistics:
  - Date, start time, stop time
  - Numbers of IPv4, IPv6, unknown packets
  - Transmission rate in pkts/s, bits/s
  - Link utilization (%)
  - Average packet size & graph of packet size distribution
  - Graph of packet size distributions (for IPv4 and IPv6)

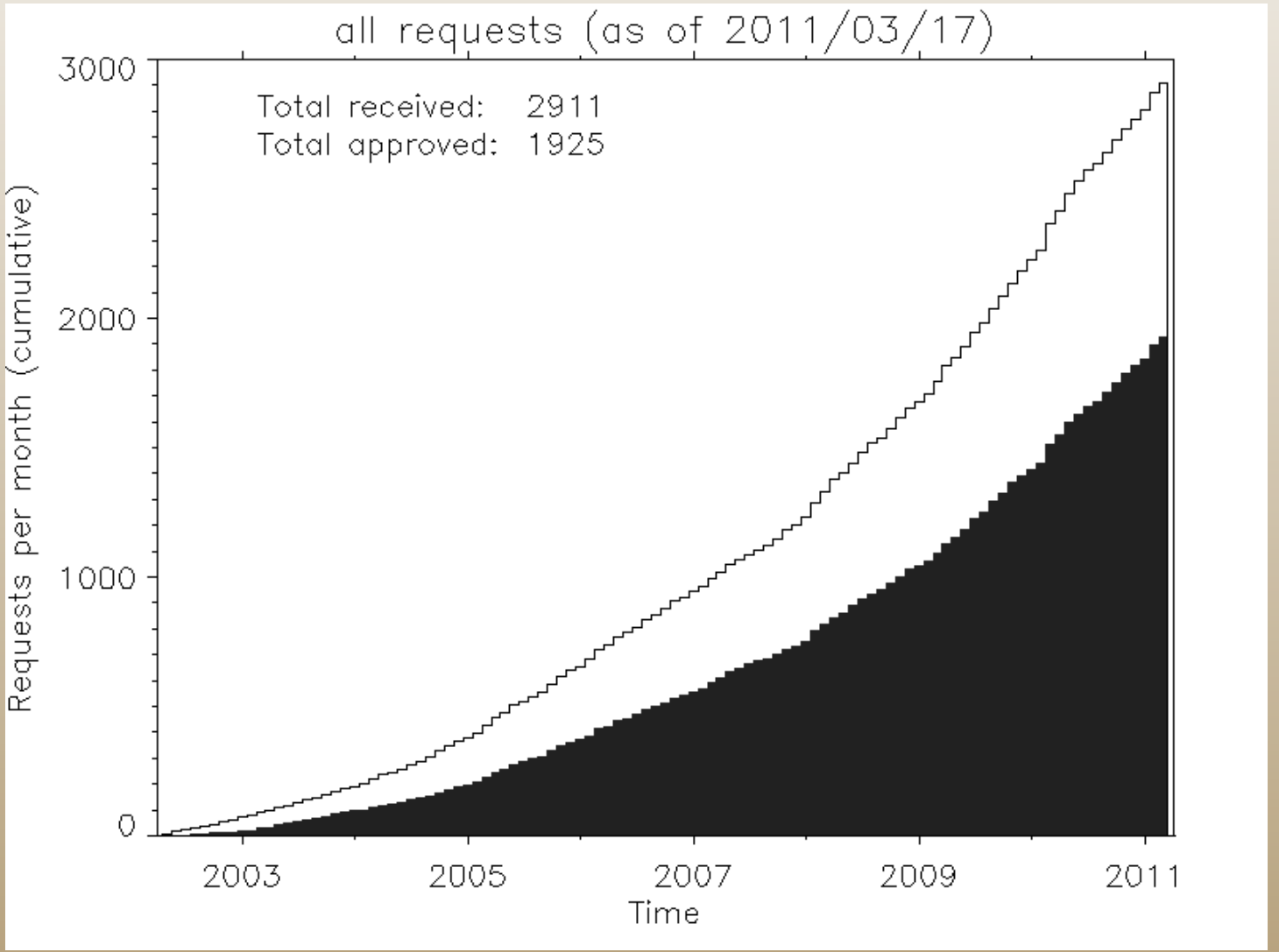
http://www.caida.org/data/passive/trace_stats/

Requests for the data, 2011/2010/2009

Dataset	Requests	Approved	Accessed	Served since
Backscatter	12/5/101	8/34/62	5/23/45	Feb 2003
Passive	39/148/242	32/115/181	55/99/151	Feb 2004
Topology	36/144/136	26/83/90	44/55/63	Jul 2004
Witty	5/13/28	4/11/18	4/10/14	Mar 2008
Telescope	3/25/35	2/19/20	2/15/16	Jul 2009
DNS-RTT	0/5/7	0/3/3	0/2/3	Aug 2006
	95/389/549	72/265/376	110/204/292	

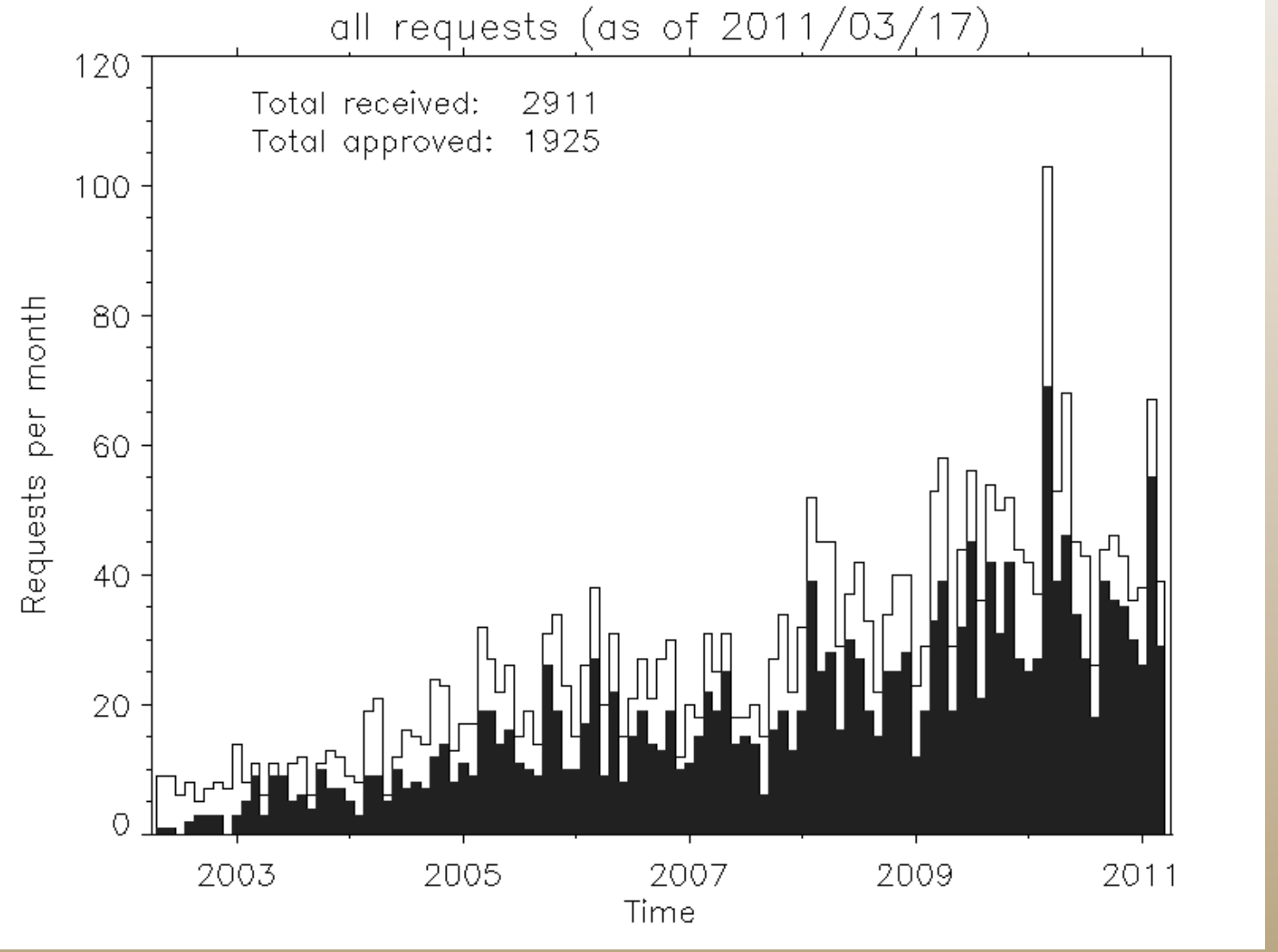
Data request stats

- all requests cumulative



Data request stats (cont)

- All requests (monthly)



Data Set Popularity (2009jan-2011mar)

1st best - OC192 and OC48 traces

- popularity: 477 requests, 378 approved, 322 accesses
- who: 219 .edu, 110 .cn, 39 .uk, 26 .com (2009/2010/2011)
  - 48 more domains

2nd best - topology data

- popularity: 332 requests, 230 approved, 164 accesses
- who: 226 .edu, 101 .cn, 36 .uk, 26 .kr, 23 .jp, 24 .com
  - 45 more domains

Publications using CAIDA data

- OC192 and OC48 traces: traffic classification, performance modeling, monitoring, filtering, generation, locality
  - http://www.caida.org/data/publications/bydataset/index.xml#passive
    - 76 publications (54 from data in PREDICT)
- UCSD telescope: Conficker, worm research
  - http://www.caida.org/data/publications/bydataset/index.xml#Backscatter
    - 26 publications (all from data in PREDICT)
- topology: pkt traceback, marking, DOS defense, topo and routing modeling, discovery, metrics, improvements
  - http://www.caida.org/data/publications/bydataset/index.xml#Topology
    - 57 publications (45 from data in PREDICT)

Recent publications

- P. Merindol, B. Donnet, J.J. Pansiot, M. Luckie and Y. Hyun, MERLIN: MEasure the Router Level of the INternet, Universite catholique de Louvain, TR 2010-3, September 2010 (Tech. Report).
- A. Dianotti and kc claffy, Obstacles and challenges to traffic classification, accepted to IEEE Network (Jan 2012?)..
- Menlo next draft (core and companion)...

Phase II Data Sets

- Provided data set descriptions for:
- UCSD telescope: near real time
- topology: Ark data (ongoing)
  - IPv4 Routed /24 Topology dataset
  - IPv4 Routed /24 DNS Names dataset
  - IPv6 Routed Topology dataset
- topology: updated ITDK 2010
- OC192 backbone: 2007-10
  - Served by CAIDA
  - Recently completed: new MOC to cover this data
    http://www.caida.org/data/collection/aup/internet_traffic_collection_moc.xml

CAIDA Master AUP

- Mission: create a master AUP
- 4 data categories (and levels of sensitivity)
  - real-time telescope data
  - passive traces
  - active traces
  - derived topology
- Uncontrolled proliferation
  - 7 data request forms
  - 22 data set web pages
  - 22 README files
- Status: First draft of master AUP (beta)
  - Will share next draft before next meeting

(reminder)

Analysis of CAIDA AUPs

- Access conditions
  - Accreditation, validation, transparency
- Use restriction
  - Purpose, probing, other
- Disclosure obligations
  - Publication, 3rd party transfer, attribution
- Enforcement
  - Compliance, attestation
- Corrections / amendments
  - Measurement error notifications
- Disposition
  - Account closure, renewal
- Policy Vehicle: AUP, MOA, MOC…

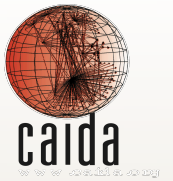
Policy Tech Transfer

- RIPE-NCC making use of PS2 framework and EIA materials to help create a guide for its users.
- Our RIPE-NCC liaison, Emile Aben, working with RIPE-NCC legal personnel (Kate) and consulting with Erin Kenneally.

Feedback on BotNet Legal Guide

- Advise to reframe as “enabling” rather than “prohibiting”
- Emphasis should be reversed from “if spam was intercepted, content may not be disclosed or used other than by provider” to “how to do relevant research while in compliance with existing laws.”
- To serve as a tool it must be a fraction of its current size.
- Detailed feedback sent to PI list

Researcher experience with PREDICT



- Case study
- Wanted to document first-hand experience with researcher data request.
- PREDICT offered a data set we could use.
- CAIDA Researcher Bradley Huffaker applied for PREDICT data.
- Next 3 slides summarize our lessons learned

Notes and Suggestions re: Application

- Make ‘Join the Community’ label more explicit, descriptive, and prominent.
  - Not at all clear how to get an account
- Restructure ‘Learn More’ page
- Simplify Account Request Form
  - Remove ‘Authorized Representative Details’ from initial request form
  - Remove password requirement
- Password Standards page too long and redundant

Notes re: Provider-Researcher MOA

- General Problems
  a) primarily prohibitive
  b) too long
  c) between Provider & Researcher or PCC & Researcher? text is inconsistent
- Specific Problems
  a) constricts even research use of data too narrowly:uses which are not permitted are explicitly denied; presumes that non-described use is always improper.
  b) restrictions on release of derivative work are too strong, under “might allow for the identification/deduction” standard.
  c) “default-deny” approach to publications is not conducive to productive research use of the data.
  d) requirement to seek review of publications by the PCC/PRB continues after the MOA expires, has no end date?
  e) requirement to “destroy all copies of the data” cannot be reasonably fulfilled. backups? destroy all derivative files? sign the MOA, yet data destruction requirements are only specified at expiration?

Notes re: Provider-Researcher MOA

- Minor Problems
  - needs to be changed if international users are to be allowed
  - does a DP always “grant a license” to a Researcher? (explicit between R and PCC, but not DP and PCC.)
  - if an individual Researcher moves, why should remaining researchers listed in the same application lose their access?
  - attribution to PREDICT as required in MOA is inaccurate, PREDICT is not “the source” of data.
  - Confidentiality Agreement attached to the MOA was for a different kind of data than Brad requested.
- Overall impression
  - NOT research friendly, NOT user friendly

CAIDA Marketing Efforts

- CAIDA web site
  - Annual reports, Program Plan, Project web page
  - will blog about Phase II
- Presentations
- Publications
- Proposals
  - NSF SDCI - open meta-data research, with expanded relevance to cyber security
  - BAA-11-02 mapping proposal: will use PREDICT
- Connections
  - Opportunity to increase synergy with NSF in new Data Management Policy
- CAIDA workshops

Necessary conditions of success

- Convenience
- Marketing
- Regular updates with *recent* data

- Phase II must be evaluated with regard to these conditions.
- Needs discussion

Open issues for Phase II

- Improve the Portal - both “how it looks” and “how it works”
  - Paul Hick worked with RTI as a portal tester; gave feedback
  - Who else participated?
  - Can we see aggregation of suggestions?
- Revise meta-data to be made public
- List of keywords - where? Or when?
- MOA revisions - we will need time!
  - 30 days to provide first revisions, 30 days for iterations
- Status of DP & DH MoAs
  - we received drafts of revised DP & DH MoAs 11 January 2011
  - were asked for review and comments within 3 days
  - emailed concerns regarding lack of sufficient time
  - 60 days still required

Open issues for Phase II (cont.)

- How to organize meta-data? - not easy!
  - how many data sets? tens? hundreds?
  - presentation
  - hierarchy
  - scalability
  - searchability
- Years of experience w/ cataloging metadata
  - huge technical and design issues in expanding catalog
  - PIs should give *early* feedback on proposed solutions
  - Avoid potentially wasted time due to poor designs
- Data categories descriptions
  - next slide.

Open issues for Phase II (cont.)

- Standardize data categories descriptions
  - Who is responsible for coherence of data descriptions?
- Proposed template
  - Essence of the category and general description of dataset
    - What aspect of the Internet is captured in dataset
    - What makes the category distinct?
    - What is scope of the data included?
  - What research questions could data be used to answer?
  - Uniform treatments, e.g., what is anonymized (not how)
  - Exemplars: special characteristics of one or more datasets?
- Did all agree to use the template? Who will enforce conformity?
- Will we discuss template for *dataset long descriptions*? Do we need them?

Other Open Issues

- Internet Topology Data Category Description
  - Submitted February 3, 2011. Portal still not updated
- Metrics to track progress
  - wiki to track PREDICT activity?
- PREDICT marketing materials
  - did 1-pager get updated
  - any other marketing materials?
- Privacy Impact Assessment update?
- Policy Section for the Portal?